Determining the accuracy in image supervised classification problems

Daniel Gómez[1] **Javier Montero**[2]

[1]Complutense University, Escuela Universitaria de Estadística
[2]Complutense University, Facultad de Matemáticas

Abstract

A large number of accuracy measures for crisp supervised classification have been developed in supervised image classification literature. Overall accuracy, Kappa index, Kappa location, Kappa histo and user accuracy are some well-known examples. In this work, we will extend and analyze some of these measures in a fuzzy framework to be able to measure the goodness of a given classifier in a supervised fuzzy classification system with fuzzy reference data. In addition with this, the measures here defined also take into account the preferences of the decision maker in order to differentiate some errors that must not be considered equal in the classification process.

Keywords: Fuzzy image classification, Accuracy measures; Kappa Index.

1. Introduction

Any supervised classification does not complete until an assessment of its accuracy has been performed. In supervised crisp image classification models it is supposed that there exist a priori knowledge (almost for a subset of items) about the belongingness to the different classes under study. In this paper, we will refer to this a priori information as the expert information. In this situations, the common way to measure the accuracy of a crisp classifier is done by selecting a sample of the data (reference data) and comparing the information (output) given by the classifier and by the expert. This information is compared generally using an error matrix. From this error matrix, it can be found in literature too many accuracy measures as overall, kappa [11], weighted kappa [12] among others that can be used (see [14] for a general review of different accuracy measures use in remote sensing image classification).

We can conclude that even there is still an open problem how to asses correctly the accuracy in supervised crisp image classification models, the problem has been studied for many researches and for real situations it can be done corrects approximations. Unfortunately, there exist other situations in which a crisp classification is not successful to represent the reality.

"Soft" classification models are desirable in many situations (specially in image processing see [23]). For example, in a probabilistic framework, the classifier could be interpreted as the likelihood that the patters lie in any of a set of possible classes. Another interesting situation of soft classification is founded in the fuzzy classification problems. Fuzzy sets theory appears in a natural way as an interesting and necessary tool to model the uncertainty due to the ambiguity and/or vagueness ([31]). In this framework, the fuzzy classifier could be understood as the partial belongingness to several categories at the same time.

Although the capacity, successful and utility of "soft" classification models have been clearly proved for many researches during the last four decades, few efforts have been dedicated to define adequate methods for the evaluation of the accuracy of their outputs. Some recent works have sought to develop approaches founded on fuzzy sets, still based on the confusion matrix. Although various suggestions have been made for the evaluation of soft classifications in general, (see [8, 10, 13, 16, 19, 22, 25] among others), the fuzzy error matrix defined in [8] is one of most used approaches, as it represents a generalization (grounded on the fuzzy set theory) of the traditional confusion matrix. In this work a generalization of the error matrix was presented in a fuzzy framework as follow:

First at all and for each object p and each cell (i, j), they determined (based on the min operator) the degree to which p has been classified in class j by the expert and in class i by the classifier. After that, this information is aggregated for each object p to obtain the *fuzzy error matrix* (see [8] for more details).

Although this fuzzy error matrix presents some clear advantages compared with the classical approaches, some misbehavior appears when the fuzzy classification that is evaluated is not a Ruspini partition [26]. Some authors (see for example [28, 17]) have partially analyzed and studied this problematic. With the main aim to fix some of these problems, in [17] it was defined a new family of disagreement weighted measures that permits to extend the most popular accuracy measures in image supervised classification: overall and the Kappa statistic for classical *hard* (crisp) classifications considering not only the main diagonal of the fuzzy error matrix

as is done in [8].

Nevertheless, the use of the Kappa index to validate the results in image classification has been recently critiqued (see [24] for more details) due to the confusion between similarity in quantity and similarity of location. In [24] it is introduced two statistics to separately consider similarity of location and similarity of quantity.

In this work, and based on the fuzzy error defined in [17], it is extended these two concepts for soft classifications. The present paper is organized as follows: in section 2, we review the main classical accuracy measures for the crisp case (overall, kappa index, Kappa location, Kappa histo and the Kno accuracy index). In section 3 the family of disagreement measures for fuzzy classification proposed in [17] is presented. In section 4, we extend and analyze the classical accuracy measures defined only for the crisp case. Finally, in section 5 some remarks and comments are drawn.

2. Accuracy assessment in crisp images classification models

As pointed out in [29], the accuracy assessment of a supervised image classification problem involves three different steps: the sampling design, the response or measurement design to obtain the true classes for each sampling (usually requiring an expert) and the analysis of the obtained data. In this work we will focus on the third step in which it is compared the results obtained by the classifier with the reference data set. The methodologies for this comparison can be divided (see [10] for more details), between *non-site specific assessments* [25] and *site specific assessments* [11]. In a non-site specific assessment, only total areas for each category are computed and compared. This approach is less expensive, but has been criticized because it does not take into account the correct localization of the classifier. Given the limitations of non-site specific assessment the site specific assessment is usually preferred. To this reason, in this work we will focus on the site specific assessment.

The common way to measure the accuracy in the site specific assessment, is by using the error matrix. The error matrix is a table that displays statistics for assessing supervised classification accuracy by showing the degree of misclassification among classes. The error matrix is also known as a confusion matrix, a contingency table or a classified error matrix. From now on, we will denote by n_{ij} the number of items that have been classified by the expert into the class j and by classifier in the class i. One one hand, we will denote by $n_{i.} = \sum_j n_{ij}$, the number of items that have been classified into the class i and by $n_{.j} = \sum_i n_{ij}$ the number of items that have been classified by the expert into the class j. We will denote by n the number of items that have to be classified.

Once the error matrix has been built, results can be compared by existing statistical techniques (non-site specific assessment techniques). One of the most popular measures obtained from the error matrix is the overall accuracy. Overall accuracy evaluates the percentage of cases correctly classified, so its interpretation is direct. Although the overall accuracy measure is really easy to interpret, we can find too many references in which this measure has been criticized. Among other things, there are some cases in which the correct classification is made by chance. In order to solve this problem, Cohen [11] defined the most widely used statistic for the estimation of the effect of change agreement, called the Kappa statistic.

Let us consider a fixed an image, divided into a set of pixels P, with $T \subset P$ the family of pixels to be tested. Let $A_1, \dots A_k$ be the set of crisp classes under consideration. The error matrix N is then defined as a frequency matrix, where each element n_{ij} represents the number of pixels that the expert classified a pixel in A_i but the classifier did in A_j.

DEFINITION 2.1 *Given the error matrix $N = (n_{ij})$, the overall accuracy is defined as*

$$O^c = \frac{\sum_{i=1}^{k} n_{ii}}{|T|}$$

being $|T|$ the number of pixels we are testing.

DEFINITION 2.2 *Given the error matrix $N = (n_{ij})$ the Kappa statistic is defined as:*

$$K = \frac{O^c - p_e}{1 - p_e}$$

where p_e represent the percentage of items that has been classified correctly by change, that is: $p_e = \frac{\sum_i n_{i.} n_{.i}}{n}$.

Again, although the Kappa statistic is the most popular technique for comparing different classifiers, it has been also extensively criticized. Some authors, see for example [15, 24], point out that the Kappa coefficient is not the only way to compensate for chance agreement or to test the significance of differences in accuracy among classifiers.

Recent studies about the Kappa index [24] permit to dissected the Kappa index into two further statistics in the framework of image classification: *Kappa location* [24] and the *Kappa histo* [20]. These two statistics are sensitive to respective differences in location and in the histogram shape of all the categories. By $Klocation$ and $Khisto$, we will denote these two statistics. In this framework is also defined the Kno agreement index [24], as the standard kappa index in which the probability of change p_e is now calculated as $\frac{1}{k}$, where k is the number of classes.

DEFINITION 2.3 *Given the error matrix $N = (n_{ij})$ the Kappa Location is defined as:*

$$Klocation = \frac{O_c - p_e}{MQPL - p_e}$$

where $MQPL$ represents the accuracy assessment situations in which the ability to specify the quantity is medium and the ability to specify the location, that is: $MQPL = \frac{\sum_i min\{n_{i.}, n_{.i}\}}{n}$ instead of 1 as in the original Kappa index.

The $Klocation$ gives the similarity or agreement scaled to the maximum similarity that can be reached with the given quantities. An alternative expression for the agreement of the quantitative model results is the maximal similarity that can be found based upon the total number of cells taken in by each category. This is called $MQPL$. $MQPL$ can be put in the context of Kappa and Klocation by scaling it to O_c. The resulting statistic is newly introduced here and is called $Khisto$, because it is a statistic that can be calculated directly from the histograms of two maps.

DEFINITION 2.4 *Given the error matrix $N = (n_{ij})$ the Khisto is defined as:*

$$Khisto = \frac{MQPL - p_e}{1 - p_e}$$

The definition of $Khisto$ has the powerful property that Kappa is now defined as the product of two factors (i.e. $K = Klocation\ x\ Khisto$). The first factor is $Klocation$, which is a measure for the similarity of spatial allocation of categories of the two compared maps. The second factor is $Khisto$, which is a measure for the quantitative similarity of the two compared maps.

Finally, considering that the probability of change p_e can be estimated as $\frac{1}{k}$, where k is the number of classes. The Kno is defined as follows:

DEFINITION 2.5 *Given the error matrix $N = (n_{ij})$ the Kno agreement index is defined as:*

$$Kno = \frac{O_c - \frac{1}{k}}{1 - \frac{1}{k}}$$

EXAMPLE 1 *Let us consider the example given in [10],where three crisp classes A_1, A_2 and A_3 were considered: Forest, Wetland and the Urban Areas, respectively. If the error matrix is*

$$M = \begin{pmatrix} 23 & 9 & 6 \\ 3 & 18 & 5 \\ 4 & 3 & 29 \end{pmatrix}$$

is easy to obtain the overall accuracy $O_c = 0.7$, the Kappa Index $K = \frac{0.7-0.336}{1-0.336} = 0.548$, the Kappa location $Klocation = \frac{0.7-0.336}{0.92-0.336} = 0.623$, the Kappa histo $Khisto = \frac{0.92-0.336}{1-0.336} = 0.879$ and the Kno index $Kno = \frac{0.7-0.333}{1-0.333} = 0.55$.

Another important topic in accuracy assessment is the importance-equivalence of the errors. In the previous approaches, all errors have been assumed of equivalent importance. This is, of course, an unrealistic hypothesis. But weights can be introduced in order to account for differences among the errors.

In this sense, the *weighted Kappa* defined by Cohen [11] incorporates unequal error weights. But, as discussed in [10], the weighted Kappa has not received too much attention. One of the reasons for this is the difficulty for the expert of correctly determining the weights (see [18] for a possible solution).

DEFINITION 2.6 *Given the error matrix $N = (n_{ij})$, the weighted Kappa statistic is defined as:*

$$K = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} u_{ij}\frac{n_{ij}}{n} - \sum_{i=1}^{k}\sum_{j=1}^{k} u_{ij}\frac{n_{i+}}{n}\frac{n_{+i}}{n}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} u_{ij}\frac{n_{i.}}{n}\frac{n_{.i}}{n}}.$$

We shall obviously assume that $u_{ii} = 1$ for all i).

3. Measuring the errors in "soft classification models

Traditional image accuracy assessment assumes crisp classes, in such a way that agreement between the classifier (C) and the expert (E) is modelled according to a two-valued model: perfect agreement (0) or total disagreement (1). This restriction implies a strong oversimplification of reality, since the continuum of variation in many landscapes will be difficult to be properly represented. In order to address this issue, we will propose a continuous error measure that summarizes the differences between a crisp reference data set (most expert are still crisp) and a fuzzy classifier.

From a mathematical point of view, a pixel being classified by the expert (E) or by the classifier (C) as the crisp class A_i, can be modelled as a k dimensional vector, k being the number of different classes under consideration, in such a way that all coordinates take value 0 except the i-th coordinate, which takes value 1. A crisp classifier or a crisp reference data set can be then considered as a function assigning to each pixel p a vector in

$$\left\{ x \in \{0,1\}^k / \sum_i x_i = 1 \right\}$$

Hence, in case our k classes are fuzzy and we assume that assignments is made in terms of a Ruspini's partition [26] (see also [6]), both classifier C and expert E will be defined as mappings

$$E : P \longrightarrow \left\{ x \in [0,1]^k / \sum_i x_i = 1 \right\}$$

and

$$C : P \longrightarrow \left\{ x \in [0,1]^k / \sum_i x_i = 1 \right\}$$

Following this notation, disagreement between classifier and expert can be measured by means of a distance in such a k dimensional space, assigning to each pixel a real value

$$D : T \longrightarrow \Re$$

where T is the subset of pixels (or polygons) selected for the accuracy assessment (an alternative valuation set can be given in terms of linguistic terms, see [16]). A standard measure for disagreement between classifier and expert is given below.

DEFINITION 3.1 *Given an image P and a family of classes $A_1 \dots A_k$ under consideration, E an expert function and C a classifier function, then the error of the pixel p given by the classifier C with reference data set E is defined as:*

$$D_f(E(p), C(p), p) = \sum_{i=1}^{k} |E(p)_i - C(p)_i|$$

where $E(p)_i$ is the i-th coordinate of $E(p)$ and $C(p)_i$ is the i-th coordinate of the $C(p)$.

Notice that this definition does not assume a crisp expert neither a crisp classifier. If classes are fuzzy in nature, mathematical models should acknowledge such a situation, and the expert should give information in terms of fuzzy classes. But in practice we find that most expert classifiers are crisp, perhaps because of the complexity in defining all parameters of a fuzzy classification. In this case, the above distance is unrealistic, giving excessive distance values to transition zones. To the aim to represent the disagreement in a more realistic way, in [17], it was proposed the following definition.

DEFINITION 3.2 *Given a an image P and a family of classes $A_1 \dots A_k$ under consideration, E a crisp expert function and C a fuzzy classifier function, then the error of the pixel p given by the classifier C with reference data set E is defined as:*

$$D_c(E(p), C(p), p) = \sum_{i=1}^{k} w_{ij} |E(p)_i - C(p)_i|$$

where $E(p)_i$ is the i-th coordinate of $E(p)$, $C(p)_i$ is the i-th coordinate of the $C(p)$, j is represents the class to which p is assigned ($E(p)_j = 1$) and each $w_{ij} \in \Re^+$ represents the importance of the error when a unit sampling that belongs to class j is classified as class i.

Notice that whenever both classifier and reference data are crisp, the above error function can be viewed as a weighted error function that takes value w_{ij} if the expert $E(p)$ has classified the pixel as A_j and the classifier has classified the same pixel as A_i. Moreover, if we take $w_{ij} = 1, \forall j$, then the disagreement measure is just the classical one. In a more general approach, these weights can depend on the maximum value E takes, or any other dispersion measure for E.

Notice also that our approach does not impose any particular structure on the classification system (as pointed out in [5, 6], fuzzy partitions in the sense of Ruspini as quite often unrealistic).

4. Accuracy measures in soft image classification

Once the error (agreement) function is obtained and the weights are determined, it is possible to define the overall accuracy, the kappa index, $Klocation$, $Khisto$ and Kno can be obtained aggregating in an adequate way the error for each object.

DEFINITION 4.1 *Given P the object set, $A_1 \dots A_k$ the family of classes under consideration, E the expert function and C the classifier function, we define the overall accuracy (O^C) as:*

$$O^C = \sum_{p \in T} \frac{1 - D(E, C, p)}{|T|} = \sum_{p \in T} \frac{A(E, C, p)}{|T|}$$

REMARK 4.1 *Let us note that if the classifier produces a Ruspini's partition (i.e., $\sum_{i=1}^{k} C_i(p) = 1$, $\forall p \in T$), and the expert is crisp, then the overall accuracy measure above defined coincides with the overall accuracy defined in [8].*

For a more general case, in case Ruspini's assumption is not fulfilled, Binaghi's approach may produce strange results, as shown below.

EXAMPLE 2 *Let us suppose that all errors are considered equally important, $E(p) = (0.2, 0.1, 0.2)$ and $C(p) = (0.4, 0.3, 0.2)$ for an object $p \in T$. If we build the fuzzy error matrix defined in [8], we obtain:*

$$X = \begin{pmatrix} 0.2 & 0.2 & 0.2 \\ 0.1 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 \end{pmatrix}$$

In this matrix, $x_{ij} = min\{E(p)_i, C(p)_j\}$. Following this fuzzy error matrix, the overall accuracy defined in [8] is $\frac{\sum_i x_{ii}}{\sum_i E(p)_i} = \frac{0.2+0.1+0.2}{0.2+0.1+0.2} = 1$. So, Binaghi's approach suggests a perfect agreement between the expert and the classifier. But in our opinion this is inappropriate. On the contrary, our agreement measure suggests a more accurate difference between expert and classifier: $A(E, C, p) = 0.8$.

EXAMPLE 3 *Let us suppose that all errors are considered equally important, $E_1(p) = (0.4, 0.1, 0.2, 0.3)$, $E_2(p) = (0.4, 0.3, 0, 0)$, $E_3(p) = (0.4, 0, 0.1, 0.3)$, $E_4(p) = (0.4, 0, 0.2, 0)$ and*

$C(p) = (1, 0, 0, 0)$ *for a given object* $p \in T$. *In the Binaghi case, the overall accuracy can be assigned an overall accuracy of* 1 *in all four cases. But again this result is not appropriate, and in fact our agreement measure establishes differences between expert and classifier:* $A(E_1, C, p) = 0.4$, $A(E_2, C, p) = 0.7$, $A(E_3, C, p) = 0.6$ *and* $A(E_4, C, p) = 0.8$.

Now an *Extended Kappa statistic* (that we will denote as K_E) is proposed, based on the previous Kappa statistic but allowing comparisons between arbitrary classifiers (a crisp classifier with a crisp data reference set and equal weights, a crisp classifier with a crisp data reference set and non-equal weights, a fuzzy classifier with a crisp data reference set and equal weights, and a fuzzy classifier with a crisp data reference set and non-equal weights).

DEFINITION 4.2 *Given* P *the object set,* $T \subset P$ *the accuracy data set with cardinality* t, $A_1 \dots A_k$ *the family of classes under consideration,* E *the expert function and* C *the classifier function, we define the Extended Kappa statistic* K_E *as:*

$$K_E = \frac{\hat{p_o} - \hat{p_c}}{1 - \hat{p_c}}$$

where $\hat{p_o} = O^C$ *is the overall accuracy*

$$\hat{p_c} = \sum_{p \in T} \sum_{q \in T} \frac{1 - D\left(C(p), E(q)\right)}{t^2}$$

where

$$D\left(C(p), E(q)\right) = Min \left\{ 1, \sum_{j=1}^{k} w_{ij} \left| \left(C(p)\right)_j - \left(E(q)\right)_j \right| \right\}$$

with $Max\left\{\left(E(q)\right)_{1 \le r \le k}\right\} = \left(E(q)\right)_i$.

It is important to note that this new definition is an extension of the standard Kappa measure for two raters. As happen in the classical case with the standard Kappa accuracy index, is possible to decompose in a multiplicative way the extended Kappa index by means of the extended Kappa location index and the extended histo index as we show below.

DEFINITION 4.3 *Given* P *the object set,* $T \subset P$ *the accuracy data set with cardinality* t, $A_1 \dots A_k$ *the family of classes under consideration,* E *the expert function and* C *the classifier function, we define the Extended Kappa location statistic* $Klocation_E$ *as:*

$$Klocation_E = \frac{\hat{p_o} - \hat{p_c}}{MQPL - \hat{p_c}}$$

$$where\ MQPL = \sum_i \frac{1}{|T|} Min \left\{ \sum_{p \in T} C_i(p), \sum_{p \in T} E_i(p) \right\}.$$

Let us observe that the definition of $MQPL$ is the natural definition of the $MPQL$ given in [24] in a fuzzy framework.

DEFINITION 4.4 *Given* P *the object set,* $T \subset P$ *the accuracy data set with cardinality* t, $A_1 \dots A_k$ *the family of classes under consideration,* E *the expert function and* C *the classifier function, we define the Extended Kappa histo statistic* $Khisto_E$ *as:*

$$Khisto_E = \frac{MQPL - \hat{p_c}}{1 - \hat{p_c}}.$$

Let us observe from previous two definitions that as happen in the classical measures the following equation holds.

$$K_E = Klocation_E \ x \ Klocation_E.$$

Finally, to conclude the extensions defined in this paper, an *Extended Kno agreement index* is proposed, based on the previous Kno accuracy index but allowing comparisons between arbitrary classifiers.

DEFINITION 4.5 *Given* P *the object set,* $T \subset P$ *the accuracy data set with cardinality* t, $A_1 \dots A_k$ *the family of classes under consideration,* E *the expert function and* C *the classifier function, we define the Extended* Kno_E *accuracy index as:*

$$Kno_E = \frac{\hat{p_o} - \frac{1}{k}}{1 - \frac{1}{k}}.$$

REMARK 4.2 *Let us observe that three previous accuracy measures are extensions of different accuracy measures that have been studied in literature fixing some of the problems that present classical accuracy measures in more general cases. In particular:*

- *For the case in which the classifier and the expert are both crisp and all the errors are equally important, it can be easily proved that previous accuracy index coincide with the classical Kappa index, Kappa location, Kappa histo, and the Kno accuracy index.*
- *For the case in which the classifier and the expert are crisp and all errors are not equally important, the extended Kappa coincides with the weighted kappa and the new Kappa location, Kappa histo and Kno accuracy index could be addressed as a weighted Kappa location, weighted Kappa histo and a weighted Kno accuracy index respectively permitting to represent real classification situations in which the errors must not consider equals.*
- *For the case in which the classifier is fuzzy, the expert is crisp and all errors are equally important, the extended Kappa coincides with the one defined in [8]. In [17], it can be seen some examples of a bad performance of the Kappa index defined in [8] for more general situations (especially when the expert is not a Ruspini partition).*

Pixel	Binaghi Agreement	$A(p, C, E)$	Crisp
p_1	0.8	0.8	1
p_2	0.6	0.6	1
p_3	0.9	0.7	0.5
p_4	1	0.8	0.5

Table 1: Agreement between expert and classifier in [8], $A(p, C, E)$ and after a crisp transformation.

- *For the case in which the classifier is fuzzy, the expert is crisp and all errors are equally important, the new extended Kappa location, Kappa histo and Kno accuracy index are the natural extension of the classical Kappa location, Kappa histo and Kno accuracy index using the matrix error defined in [8].*
- *For the case in which the classifier is fuzzy, the expert is crisp and all errors are not equally important, the extended Kappa could be addressed as a weighted fuzzy kappa index defined in [8] considering now different importance for the errors.*

In the following example we calculate the new extended accuracy measures comparing with the one defined in [8] and the classical crisp measures for a very simple example.

EXAMPLE 4 *Let us consider an image T with three classes A_1, A_2 and A_3. Let us suppose that the image only contain four different pixels p_1, p_2, p_3 and p_4 with frequency* 10, 20, 20 *and* 50 *respectively. The outputs of the fuzzy classifier for these four class of pixels are $C(p_1) = (0.8, 0.1, 0, 1)$, $C(p_2) = (0.6, 0.4, 0)$, $C(p_3) = (0.4, 0.5, 0, 1)$ and $C(p_4) = (0.4, 0.4, 0.3)$. The expert opinion for these four pixels are $E(p_1) = (1, 0, 0)$, $E(p_2) = (1, 0, 0)$, $E(p_3) = (0.5, 0.3, 0.2)$ and $E(p_4) = (0.4, 0.4, 0.1)$. The confusion matrix calculated by means of [8] is the following*

$$X = \begin{pmatrix} 48 & 26 & 11 \\ 37 & 26 & 9 \\ 11 & 19 & 9 \end{pmatrix}$$

and $n = 95$. The different agreement measures and accuracy measures can be viewed in tables 1 *and* 2.

First at all, we would like to remark that for the calculations of the crisp assessment in tables 1 and 2 we have transformed (as is done usually) the fuzzy information by means of the Maximum operator (i.e. The vector $(0.6, 0.4, 0)$ is transformed into $(1, 0, 0)$). For the situations in which the maximum is reached in more that one class we have adopted a probabilistic interpretation (i.e. the vector $(0.4, 0.4, 0.2)$ is transformed with probability $1/2$ into the vector $(1, 0, 0)$ and with probability $1/2$ into $(0, 1, 0)$.) Also, let us note that for the calculations of the agreement function defined in [17] we have considered all errors equally important.

Classical Overall	$\frac{65}{100} = 0.65$
Binaghi Overall	$\frac{83}{95} = 0.873$
O^c	$\frac{74}{100} = 0.74$
Classical Kappa	$\frac{0.65-0.575}{1-0.575} = 0.176$
K_E	$\frac{0.74-0.7065}{1-0.7065} = 0.1141$
Classical Kappa location	$\frac{0.74-0.575}{0.8-0.575} = 0.733$
$Klocation_E$	$\frac{0.74-0.7064}{0.83-0.7065} = 0.271$
Classical Kappa histo	$\frac{0.8-0.575}{1-0.575} = 0.529$
$Khisto_E$	$\frac{0.83-0.7065}{1-0.7065} = 0.4207$
Classical Kno	$\frac{0.65-0.333}{1-0.333} = 0.475$
Kno_E	$\frac{0.74-0.333}{1-0.333} = 0.61$

Table 2: Comparison between different Accuracy measures.

As can be observe from the previous example, the crisp assessment is unrealistic and fails due to the fact of the big information that is lost when fuzzy information is transformed to crisp information. This fact can be easily view in pixels p_1 and p_2 (perfect agreement but the opinions between expert and classifier are different) and pixels p_3 and p_4 (low agreement considering both opinions). The assessment proposed in [8] clearly improve the crisp assessment but fails in some aspects as the overestimation. This fact can be clearly viewed in the pixel p_4 (perfect agreement but the opinions between expert and classifier are different). In addition with this, the agreement between expert and classifier for the pixel p_3 is greater than the agreement for the pixel p_1 but as can be seen this situations should be inverse.

The agreement defined in [17] fix these previous situations giving a more realistic index of Overall (between Binaghi approach and crisp approach). This situations also is observed in the Kappa index, Kappa location, Kappa Histo or Kno accuracy index.

5. Final remarks

In this work, we have focused on the problem of developing accuracy measures that permit us to establish the goodness of image classification when there exist a reference data. In the crisp case, it is very common the use of statistical index as Overall. Nevertheless, the use (in the crisp case) of the overall accuracy measure to represent the quality of a classification performance has serious problem since not represent clearly the real agreement between the reference data and the classifier. To this reason the use of mainly Kappa Statistic and recently Kappa location, Kappa histo was having a great impact (especially in remote sensing) in image accuracy assessment. We would like to stress that although Kappa index can be used for any classification problem as a measure of agreement between expert and classifier, its decomposition into Kappa histo and Kappa location only has sense in the framework of image

classification (specially in remote sensing classification problem). This is the reason way the topic of this research has been focus on image classification.

Our proposal pursues an evaluation of accuracy of fuzzy classifications, extending previous accuracy measures into a fuzzy framework, following the works that were developed in [17, 18] or [8] among others. Although in this paper the accuracy assessment can be carry out in any situation (expert fuzzy, classifier fuzzy, and error not equal) we will like to emphasize that all reference data you can find in image classification literature are crisp. But as is pointed in [21] more efforts are needed in order to built fuzzy reference data sets that gather the fuzzy expert's opinions. For this case, is important to note that the disagreement measure here proposed can be easily generalized in order to be able to assess the accuracy when the classifier and the reference data be fuzzy.

References

[1] S. Haykin, editor. *Unsupervised Adaptive Filtering vol. 1: Blind Source Separation*, John Willey ans Sons, New York, 2000.

[2] N. Delfosse and P. Loubaton, Adaptibe blind separation of sources: A deflation approach, *Signal Processing*, 45:59–83, Elsevier, 1995.

[3] S. Cruces, A. Cichocki and S. Amari, The minimum entropy and cumulants based contrast functions for blind source extraction. In J. Mira and A. Prieto, editors, proceedings of the 6^{th} *international workshop on artificial neural networks* (IWANN 2001), Lecture Notes in Computer Science 2085, pages 786-793, Springer-Verlag, 2001.

[4] F. Vrins, C. Archambeau and M. Verleysen, Towards a local separation performances estimator using common ICA contrast functions? In M. Verleysen, editor, *proceedings of the* 12^{th} *European Symposium on Artificial Neural Networks* (ESANN 2004), d-side pub., pages 211–216, April 28–30, Bruges (Belgium), 2004.

[5] A. Amo, D. Gomez, J. Montero and G. Biging, Relevance and redundancy in fuzzy classification systems, *Mathware and Soft Computing*, 8: 203–216, 2001.

[6] A. Amo, J. Montero, G. Biging and V. Cutello, Fuzzy classification systems. *European Journal of Operational* 156: 459–507, Elsevier, 2004.

[7] R.G. Congalton and G. Biging, A pilot study evaluating ground reference data collection efforts for use in forestry inventory. *Photogrametic Engineering and Remote Sensing* 58: 1669–1671, 1992.

[8] E. Binaghi, P.A. Brivio, P. Ghezzi, and A. Rampini, A fuzzy set based accuracy assessment of soft classification. *Pattern Recognition Letters*, 20: 935–948, Elsevier, 1999.

[9] R.G. Congalton, A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37: 35–46, Elsevier, 1991.

[10] R.G. Congalton, and K. Green, editor. *Assessing the accuracy of remotely sensed data: Principles and practices.* London New York and Washinton D.C: Lewis publishers, 1999

[11] J. Cohen, A coeficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37–46, 1960.

[12] J. Cohen, Weighted Kappa: Nominal Scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70: 213–220, 1968.

[13] G.M. Foody, Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80: 185–201, 1995.

[14] G.M. Foody, The continuum of classification fuzziness in thematics mapping. *Photogrammetric Engineering and Remote Sensing*, 65: 443–451, 1999.

[15] G.M. Foody, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80: 185–201, 2002.

[16] S. Gopal and C.E. Woodcock, Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, 60(2): 181–188, 1994.

[17] D. Gómez, G. Biging and J. Montero, Accuracy statistics for judging soft classification. *International Journal of Remote Sensing*, 29(3): 693–709, 2008.

[18] D. Gómez, J. Montero and Biging G., Accuracy measures for fuzzy classiffication in remote sensing. B. Bouchon-Meunier and R.R. Yager, eds. (Editions E.D.K., Paris), 1556–1563, 2006.

[19] Green, K. and Congalton, G. (2004). An error matrix approach to fuzzy accuracy assessment: The NIMA geocover project. In R. S. Lunetta and J.G. Lyon (Eds.), Remote sensing and GIS accurcacy assessment (pp. 163-172). Boca Raton: CRC Press.

[20] Hagen-Zanker A. (2006). Map comparison methods that simultaneously address overlap and structure. Journal of Geographical Systems 8(2):165-185

[21] M. Laba, S.K. Gregory, J. Braden, D. Ogurcak, E. Hill, E. Fegraus, J.Fiore, and S.D. DeGloria, Conventional and Fuzzy accuracy assessment of the New York Gap Analysis Project land cover map. *Remote sensing of Enviroment*, 81: 443–455, 2002.

[22] H.G. Lewis, and M. Brown, A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22: 3223–3235, 2001.

[23] M. Nachtegael, T. Mélange, E.E. Kerre, The possibilities of fuzzy logic in image processing, Lecture Notes in Computer Science 4815 (2007), 198-208.

[24] Jr R.G. Pontius, Quantification error versus location error in comparison of categorical maps. *Photogrammetric Eng. Remote Sensing* 66: 1011–1016, 2000.

[25] Jr R.G. Pontius and M.L. Cheuk, A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science*, 20(1): 1–30, 2006.

[26] E.H. Ruspini, A new approach to clustering. *Information and Control*, 15: 22–32, 1969.

[27] T.L. Saaty (1994): *Fundamentals of Decision Making with the Analytic Hierarchy Process*. RWS Publications, Pittsburgh (Revised in 2000).

[28] J.L. Silván-Cárdenas and L. Wang, Sub-pixel confusion-uncertainty matrix for assessing soft classifications. *Remote Sensing of Environment* 112: 1081–1095, Elsevier, 2008.

[29] S.V. Stehman, and R.L. Czaplewski, Introduction to special issue on map accuracy. *Enviromental and Ecological Statistics*, 10, 301– 308, 2003.

[30] J.S. Uebersax, A generalized Kappa coefficient. *Educational and Psychological Measurement*, 42: 181-183, 1982.

[31] L.A. Zadeh, Fuzzy sets, *Information and Control* 8 338–353, 1965.